Sönning, Lukas; Grafmiller, Jason

Seeing the wood for the trees : Predictive margins for random forests

Date of secondary publication: 03.07.2024

Accepted Manuscript (Postprint), Article

Persistent identifier: urn:nbn:de:bvb:473-irb-962740

Primary publication

Sönning, Lukas; Grafmiller, Jason (2023): „Seeing the wood for the trees : Predictive margins for random forests". In: Corpus Linguistics and Linguistic Theory, Vol. 20, Nr. 1, pp. 153-181, Berlin: de Gruyter, doi: 10.1515/cllt-2022-0083.

This is an Accepted Manuscript of an article published by De Gruyter in *Corpus Linguistics and Linguistic Theory* on 28 March 2023, available at: https://doi.org/10.1515/cllt-2022-0083

Seeing the wood for the trees: Predictive margins for random forests

Lukas Sönning (University of Bamberg)
Jason Grafmiller (University of Birmingham)

Abstract. Classification trees and random forests offer a number of attractive features to corpus data analysts. However, the way in which these models are typically reported – a decision tree and/or set of variable importance scores – offers insufficient information if interest centers on the (form of) relationship between (multiple) predictors and the outcome. This paper develops predictive margins as an interpretative approach to ensemble techniques such as random forests. These are model summaries in the form of adjusted predictions, which provide a clearer picture of patterns in the data and allow us to query a model on potential non-linear associations and interactions among predictor variables. The present paper outlines the general strategy for forming predictive margins and addresses methodological issues from an explicitly (corpus) linguistic perspective. For illustration, we use data on the English genitive alternation and provide an R package and code for their implementation.

Keywords: average predictive comparisons; classification trees; interpretable machine learning; predictive modeling; random forests

1. Introduction

Recursive partitioning techniques, e.g. classification trees and random forests, have gained considerable ground in corpus data analysis. There are settings where these tools offer attractive advantages over competing methods (e.g. Strobl et al. 2009; Levshina 2020): (i) Tree-based methods are able to detect complex and non-linear relationships that may hold between (multiple) variables and the outcome; (ii) they work well on small samples with many predictors; (iii) they can handle imbalanced datasets with collinear input variables; and (iv) it is often claimed that their primary outputs, i.e. decision-tree-like representations and variable importance scores, allow for straightforward interpretation.

In this paper, we argue that while tree-and-forest techniques do offer certain advantages, the current standard for the reporting of such analyses fails to capitalize on their full potential. Showing a single "best" tree and/or set of importance scores rarely yields informative and interpretable data summaries. This is particularly true for complex data layouts with many predictor variables (Grafmiller 2023), where these tools are meant to excel. In fact, patterns are often quite difficult to read from individual trees, and interactions and non-linearities, if they exist, are not at all apparent in the standard variable importance scores (Gries 2020).

Recent work in machine learning (ML) has made some progress towards rendering the output of "black-box" models more interpretable (see Molnar 2019). Here we present a technique similar to partial dependence plots (Friedman 2001), which brings into view the patterns suggested by a random forest model. To this end, we incorporate concepts often found in the reporting of regression models, more specifically, model summaries in the form of adjusted predictions, marginal means, or marginal (or partial) effects. Following Lane and Nelder (1982: 615), we refer to these summaries as *predictive margins*, which are essentially model-based predictions for certain conditions of interest. These predicted quantities allow us to extract from a model the equivalent of what are sometimes referred to as main effects and interaction effects. Translating random forests into focused predictive margins offers a clearer picture of patterns in the data and helps us assess the degree of complexity (or lack thereof) suggested by our analysis. Further, by tapping into the ensemble of trees constituting a random forest, we are able to offer indications of statistical uncertainty, similar to confidence regions around regression-based predictions. We illustrate this technique using data on the English genitive alternation and the online supplementary materials include an R tutorial (https://osf.io/r5uca) and a custom-made R package (https://github.com/jasongraf1/predictiveMargins).

Before we go further, we should note that since prediction methods are a profoundly general strategy for understanding and interpreting statistical models, it would be naïve for us to make any claims to novelty. Instead, the main contribution of this paper is two-fold. First, we discuss this interpretative approach from a (corpus) linguistic perspective and provide substantive guidance for summarizing model-based predictions in an informative way. Second, to reach a broad audience, the text is written in a tutorial-like fashion and key concepts surrounding predictive margins are developed step by step.

The paper is structured as follows. We start with a brief review of tree-based methods and random forests. Section 3 then offers a simplified illustration of how to extract predictive margins from a random forest. In Section 4, we discuss two predictor classifications that help us coordinate our post-processing steps. This provides a foundation for Section 5, which sets out the general strategy in more detail. After a comparison of predictive margins to partial dependence plots (Section 6), Section 7 applies them to a random-forest analysis of the English genitive alternation. We contrast two data-analytic modes: A planned analysis that is driven by focused questions and a post-hoc exploration of the random forest without any fixed expectations or priorities. Section 8 concludes with a summary and outlook.

2. Recursive partitioning: Trees and forests

Like other ML tools, classification and regression trees (CARTs; Breiman et al. 1984) were developed for prediction tasks, i.e. settings where the aim is to forecast the value (or level) of an outcome variable, based on a set of input features (i.e. predictor variables). The underlying algorithm optimizes the resulting prediction scheme, the benchmark being its performance on (future) unobserved data. Tree-and-forest techniques are therefore particularly suitable for applied contexts, where the anticipation of future events can be turned to good account. In this paper, we restrict our attention to classification trees and random forests for binary outcomes.[1] The remainder of this section offers a brief overview of these tools; for more detailed accounts aimed at a linguistic audience, see Tagliamonte and Baayen (2012), Levshina (2020), and Grafmiller (2023).

[1] It should be noted that the approach we develop easily extends to nominal and quantitative outcome variables.

For illustration, Figure 1 shows a classification tree for the genitive data based on only two predictor variables, ANIMACY (animate, collective, inanimate) and NOUN TYPE (proper vs. common noun). Splits are indicated with points, which refer to the splitting variable (small caps) and the levels going into the branches. The terminal nodes show the predicted probability of the *s*-genitive. We can see that the likelihood of observing an *s*-genitive is highest for animate possessors referenced by a proper noun (.85, or 85%).

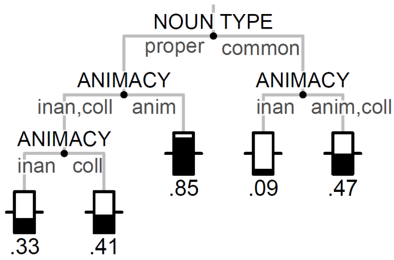

Figure 1. Illustrative classification tree for the genitive data based on only two predictors. 🅭🅯[2]

A tree can generate a prediction for each constellation of predictor values (or levels). For a binary classification task, we obtain the predicted variant and/or the predicted probability of each variant. To this end, we trace a path through the tree. At each split, the relevant predictor value tells us which branch to choose. We end up at a terminal node with a predicted probability.

To improve the performance of CARTs, various ensemble methods have been devised, among them random forests (Breiman 2001). As the name suggests, a random forest is an ensemble of trees. To run a forest of (say) 500 trees, random perturbations are introduced into the analysis. The first chance element applies to the data. Typically, each tree is constructed from a bootstrap sample of the original data table, which means that the random forest is grown on 500 slightly different sets of observations. Random forests add a second element of randomness in the selection of predictor variables that are considered at each split. For each splitting occasion, the algorithm will randomly select a subset of *m* predictors, and the best split on these *m* predictors is then used to partition the data. The procedure continues recursively, with each node involving a potentially different subset of predictors, until no further splits are justified. By contrast, a standard CART always evaluates all predictors to find an optimal split.

These two chance components produce a (potentially very) heterogeneous collection of trees, which compensates for sparse values in the data while at the same time giving greater consideration to weaker predictors and inputs "masked" as a result of collinearity. Since each tree generates a unique forecast for a particular combination of predictor values (e.g. animate possessors referenced by a proper noun), our random forest yields 500 predicted

[2] Images with the symbols 🅭🅯 in the figure caption have been published under the Creative Commons Attribution 4.0 license (CC BY 4.0, http://creativecommons.org/licenses/by/4.0) in the accompanying OSF project (https://osf.io/4tuv2/).

probabilities for a given condition. It turns out that the average over these 500 scores gives a more accurate forecast compared to a single tree run on the full set of data. Further, the variation among the 500 predicted probabilities indicates the uncertainty surrounding the averaged predictions.

Due to their explicit focus on predictive accuracy, tree-based methods are useful tools for many real-world applications. In basic science, however, the data-analytic mentality is often more strongly governed by substantive and theoretical concerns, with a focus on inference rather than prediction (but see Yarkoni and Westfall 2017). Thus, while applied research often takes a forward view, aiming to produce accurate forecasts, basic research usually takes a backward view, aiming for a better understanding of the underlying processes that generated the data. In corpus-based work, examples of applied goals include authorship attribution, the provision of forensic evidence, or building accurate annotation software. Basic-science corpus research, on the other hand, aims to explain linguistic phenomena by relating to domains such as cognition (e.g. complexity-related factors), discourse (e.g. accessibility of referents), or the physiology of articulation (e.g. economy). For the genitive alternation, for instance, we want to understand *why* a particular variant occurs, and we therefore focus our attention on predictor variables and the patterns they form. Concerns of predictive accuracy, on the other hand, may be secondary.

Typically, in tree-based methods the attention to predictor variables is limited to assessing their relative contribution to model performance (see, e.g., Tagliamonte and Baayen 2012; Levshina 2020). The status of a predictor is reflected in its importance score, an abstract metric that allows us to rank inputs in terms of their predictive ability. While these summary scores do provide an overall assessment, no information is given about predictor-specific patterns. Thus, while it is useful to know that, for the English genitive alternation, ANIMACY tops this list, we may also wonder how predictions differ across its levels. Variable importance scores and the overall predictor rankings they produce are therefore quite limited in terms of informativity.

Predictor-specific patterns can sometimes be reconstructed based on an inspection of classification trees (e.g. Szmrecsanyi et al. 2016: 118). However, even for simple trees such as that shown in Figure 1, this can be mentally taxing. Predictive margins allow us to obtain more tangible summaries, which can be tailored to address a variety of specific questions about patterns in the data.

3. Predictive margins based on random forests

This section demonstrates the basic strategy for obtaining predictive margins. To this end, we present a rough outline of the procedure using a miniature analysis of the genitive data; subsequent sections expound both the technique and the analysis in greater detail. The data, which are available via TROLLing (Grafmiller 2022), include 5,098 cases in total, which are drawn from two corpora, Brown (Francis and Kučera 1964) and Frown (Hundt et al. 1999). Analyses were conducted in R (R Core Team 2022) using the *ranger* package (Wright and Ziegler 2017) for fitting random forests. An R script for reproducing the analyses and figures in this paper is available from the associated OSF repository (https://osf.io/gvymt).

For the moment, we concentrate on three predictor variables:

- ANIMACY of the possessor: animate, collective, inanimate
- NOUN TYPE of the possessor: proper noun, common noun
- GENRE: general fiction, learned, non-fiction, press, adventure fiction

The outcome variable is the form of the genitive (*s*- or *of*-genitive). To avoid redundancy, from here on we will only report the probability of *s*-genitives.

Once we have fit a random forest (of 500 trees) to the data, predictive margins are formed in three steps. We first construct a reference grid, which defines the conditions for which we wish to obtain predictions. By condition, we mean a specific combination of predictor values. For our current data set, this grid includes all possible combinations of ANIMACY (3 levels), NOUN TYPE (2 levels), and GENRE (5 levels). This adds to 30 conditions in total. Table 1 shows an excerpt from this data grid.

Table 1. Excerpt from the reference grid with the predicted probability of the *s*-genitive from each of 500 trees in a random forest

Reference grid				Predicted probability				
Condition	ANIMACY	NOUN TYPE	GENRE	1	2	…	500	**Median**
1	animate	common	general fiction	.59	.55	…	.44	**.53**
2	animate	common	learned	.47	.44	…	.54	**.48**
3	animate	common	non-fiction	.41	.70	…	.44	**.44**
4	animate	common	press	.59	.70	…	.60	**.58**
5	animate	common	adventure fiction	.59	.52	…	.60	**.56**
6	animate	proper	general fiction	.87	.89	…	.91	**.87**
7	animate	proper	learned	.87	.90	…	.87	**.87**
8	animate	proper	non-fiction	.87	.70	…	.84	**.84**
9	animate	proper	press	.87	.86	…	.82	**.87**
10	animate	proper	adventure fiction	.87	.91	…	.91	**.90**
…	…	…	…	…	…	…	…	…

The second step is to obtain predictions for these conditions from the random forest. For each condition, our forest generates 500 predicted probabilities, one from each tree. Figure 2 shows one tree in our forest. For illustration, consider condition 2 (animate possessor, common noun, learned writing). As we trace its path in the tree, we end up at the right-most node and a predicted probability of .47 (i.e. 47%) for the *s*-genitive.

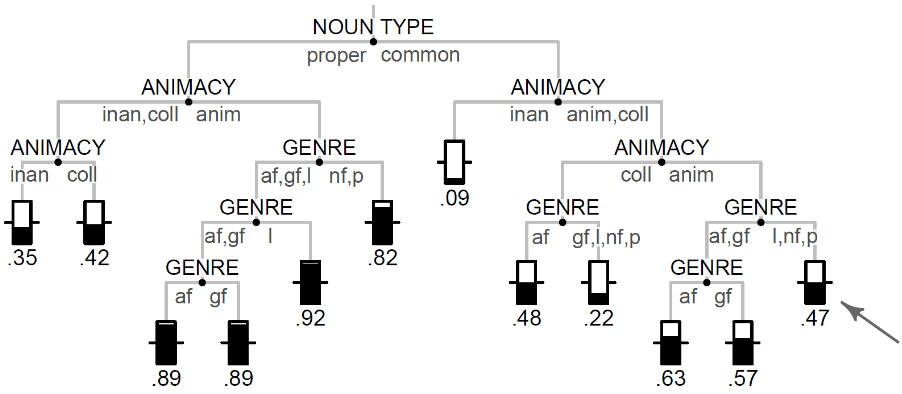

Figure 2. A classification tree from the random forest, based on three predictor variables (ANIMACY, NOUN TYPE, GENRE).

If we do the same for the other trees in the forest, we obtain 500 predicted probabilities for condition 2. Recall that, due to the perturbations induced when growing a forest (data bootstrapping and predictor sampling), the individual trees will differ from one another. As a result, they will produce heterogeneous predictions for a specific condition. The distribution of the 500 scores for condition 2 appears in Figure 3, where each dot denotes the forecast from a single tree. The value from our illustrative tree (.47) sits near the lower end of this distribution. Upon completing step 2, we end up with 30 such distributions, one per condition of interest.

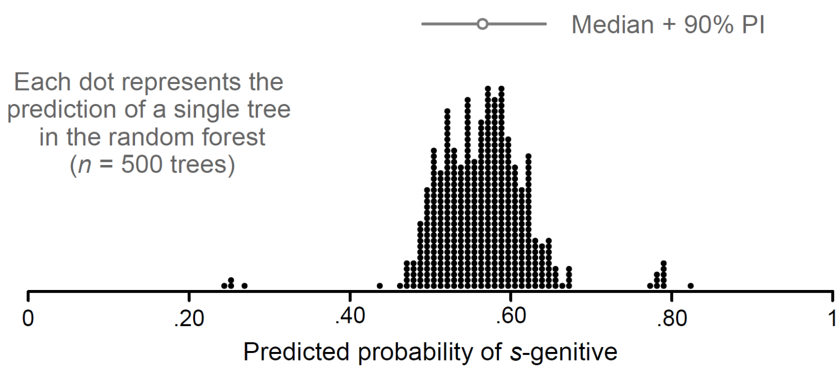

Figure 3. Probability of the *s*-genitive in condition 2 (animate, common noun in the genre 'learned'): Distribution of the 500 predicted probabilities from the random forest.

In the third step, we summarize these predictions. To keep things simple for now, we form predictive margins for two-way classifications based on the predictors ANIMACY and GENRE, while disregarding (i.e. averaging over) NOUN TYPE. This still leaves us with a set of 500 predictions each for 15 conditions of interest (i.e. all combinations of GENRE and ANIMACY). We condense these distributions to a few expressive indicators: a measure of central tendency (the median) and a measure of variability, a percentile interval (PI) that covers the central 90% of the predictions. In Figure 3, this summary (*Mdn* = .56, 90% PI [.49; .64]) appears above the dot diagram.

We can then analyze and present these predictive summaries, either in tabular or visual form. With two predictor variables, there are different ways of arranging these in a display (see Sönning 2023), and we will consider both layouts. Figure 4 shows how the ANIMACY profile (animate > collective > inanimate) varies across genres. With the exception of press, all genres show remarkably similar configurations – a hockey-stick pattern. The ANIMACY cline appears to be attenuated in press writing, where collective and inanimate possessors are much more likely to take the *s*-genitive. The predictive uncertainty limits around these profiles indicate that this deviant trend is quite stable across the random perturbations underlying the random forest.

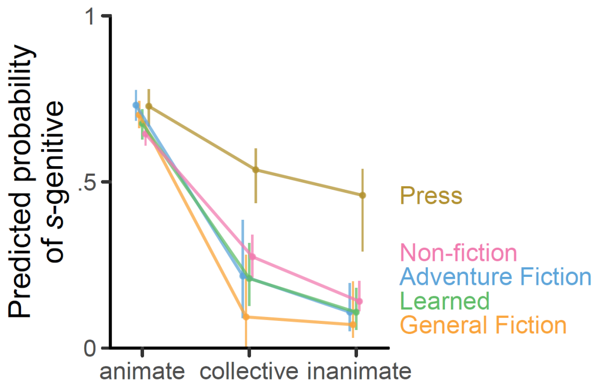

Figure 4. Predictive margins with respect to ANIMACY and GENRE: Arrangement 1. Error bars denote 90% PIs.

Figure 5 rearranges these predictive margins and puts GENRE on the horizontal axis, with categories ordered based on their overall probability of *s*-genitives. The display therefore shows three traces, one for each animacy class. What emerges from this layout is that differences among genres are greater for collective and inanimate possessors.

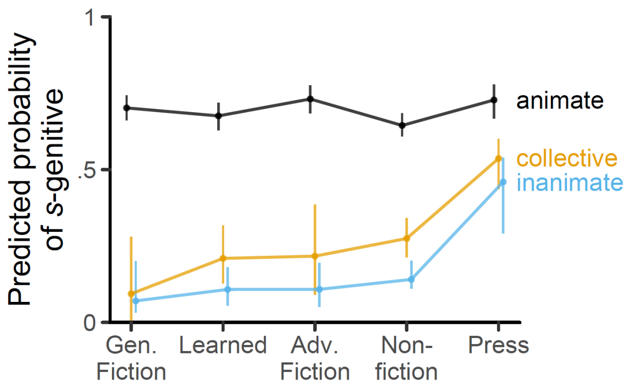

Figure 5. Predictive margins with respect to ANIMACY and GENRE: Arrangement 2. Error bars indicate 90% PIs.

To facilitate interpretation, we have simplified patterns by averaging over the levels of NOUN TYPE. This kind of aggregation requires thought, however, since averages may be formed in different ways. Thus, we may give equal emphasis to all levels, or we may wish to weight them differently, say, to reflect their actual distribution in language use. This decision primarily hinges on the nature of the predictor variables.

4. Predictor classes

When averaging probabilities to form predictive margins, the data analyst must make a number of decisions that will affect the meaning carried by these summaries. It is therefore desirable for such decisions to be made on principled grounds. To this end, the current section categorizes predictors along two dimensions, which reflect their role in the focal comparison (Section 4.1) and their relation to corpus design (Section 4.2). As we will see in Section 5, these classifications allow us to coordinate our predictive summaries.

4.1. Focal comparisons: Targeted vs. peripheral predictors

Based on the specific data-analytic question we are asking, predictors fall into two groups: targeted and peripheral variables (see Grissom and Kim 2012: 206). Targeted factors are the ones that are central to a particular comparison we intend to make. When articulating what kinds of pattern we are interested in, we make explicit reference to these variables. Since the summary of data patterns suggested by a model usually involves making different comparisons, predictors can change status during model interpretation. The distinction between targeted and peripheral variable is therefore not an attribute of the variables as such, but rather depends on the comparison of interest.

Data-analytic questions may foreground a single predictor ("Does the predicted share of s-genitives vary across animacy levels?") or multiple inputs ("Does the animacy pattern vary across genres?"). If there are multiple targeted factors, there are different ways of formulating and thinking about comparisons; it then makes sense to distinguish between primary and secondary targeted variables. These different roles are reflected in how a question is phrased. Consider the following queries:

(1) Does the animacy pattern (primary) vary across genres (secondary)?
(2) Are genre differences (primary) consistent across animacy levels (secondary)?

From a statistical perspective, the primary-secondary distinction is largely irrelevant, due to the numerical equivalence of perspectives (1) and (2). For interpretation, however, the choice between these perspectives matters since they typically produce different insights. They also map onto different graphical layouts. For instance, line plots instantiate the preferred perspective if the primary targeted variable is assigned to the horizontal axis (see Shah and Freedman 2011: 570; Sönning 2023). To appreciate this perceptual bias, note that question (1) is better answered by Figure 4, while question (2) goes hand in hand with Figure 5.

The second group of variables assumes a peripheral status for a given data-analytic query. It includes those factors that are part of the random forest, but play no role in the focal comparison(s). When aggregating predictions, we must decide on how to handle these additional variables, i.e. how to average over their levels. The basic question is whether their levels should be treated on a par, or whether we should weight them based on their distribution in the observed data. This decision is best answered on linguistic grounds, and often hinges on a distinction between internal and external variables, to which we now turn.

4.2. Corpus design: External vs. internal predictors

To recognize the basic difference between external and internal variables, consider the process of corpus compilation. When constructing a corpus, we make a number of decisions that will affect the language variety that ends up being represented. This includes, for instance, the choice of speaker groups or genres from which we sample texts. External variables, then, are attributes of the data that can be controlled or manipulated during corpus design (or data collection more generally) and reflect the purpose of the project. These features largely correspond to the variationist sociolinguistic category of social or external predictors (e.g. Labov 1972). For the genitive data, there are two external predictors: GENRE and TIME (1960s vs. 1990s, i.e. Brown vs. Frown).

Internal predictors, on the other hand, are beyond the corpus compiler's control. These attributes largely correspond to the Labovian category of internal or linguistic features (Labov 1972). They reflect the forces at work in natural language use, such as iconicity, discourse-pragmatic factors, and processing constraints. In the genitive data, most predictors are internal. Attributes of the possessor, for instance – its animacy, length, or discourse status – arise from the specifics of the communicative situation; in contrast to external features, internal variables vary naturally in the corpus.

The internal-external distinction is relevant when considering the distribution of these classes of variables in the corpus. Since internal variables reflect the natural workings of language and communication, their occurrence profile is meaningful in its own right. For instance, collective possessors are rarer (roughly 10-15% of the cases) than animate and inanimate ones (each about 40-50% of the cases). It is often desirable to preserve the distributional features of such attributes, to adjust predictions in such a way that they are reflective of actual language use.

The distribution of external variables, on the other hand, is usually not of linguistic interest. Since they reflect corpus design rather than features of natural language use, we are unlikely to attribute much meaning to them. Out of all the genres, the genre 'learned', for instance, accounts for the greatest share of the corpus (in terms of overall word count).[3] The non-fiction component, on the other hand, accounts for the greatest share of genitives (40% of the tokens in the data). For purposes of comparison, say, to another corpus, this lack of balance is usually a nuisance. To enhance comparability, we can use statistical adjustment to give equal emphasis to all genres.

Once we have categorized predictors along these two dimensions, we are set to build predictive margins in an informed way.

[3] The learned section includes 80 texts, followed by non-fiction (75), press (44) and the two fiction sections (29 each). All texts in Brown are excerpts of roughly 2,000 words.

5. Forming predictive margins

We now take a closer look at how to extract focused predictions from a random forest. We discuss and contrast different ways of averaging over the levels of peripheral predictors and close with comments on how to handle quantitative predictors.

5.1. Averaging over peripheral predictors

In settings with multiple predictor variables, it is rarely feasible to inspect the full set of conditions individually. Recall that, in our 3-factor analysis of the genitive data, we averaged over the predictor NOUN TYPE, which we had set aside as a peripheral variable, to simplify patterns for an interpretation of our targeted variables ANIMACY and GENRE. We now discuss different ways of averaging over peripheral variables.

We stick to the foregoing miniature analysis, but now consider the full set of conditions, i.e. 30 in total. The median predictions for these are given in Table 2. The probability of .59 in the top left corner, for instance, indicates that, for animate possessors referenced by a common noun in press writing, 59% of cases are predicted to be *s*-genitives.

Table 2. Median predicted probabilities of *s*-genitives by ANIMACY, GENRE, and NOUN TYPE.

	Animate		Collective		Inanimate	
GENRE	Common	Proper	Common	Proper	Common	Proper
Press	.59	.87	.49	.59	.25	.68
Adventure Fiction	.57	.90	.24	.18	.05	.15
Non-fiction	.45	.84	.24	.31	.09	.19
Learned	.48	.87	.21	.20	.04	.17
General Fiction	.53	.87	.06	.09	.07	.06

When the full set of conditions is too large to be studied comprehensively, we condense the information by averaging over certain (peripheral) predictor variables. To illustrate, let us (again) simplify our data summary and disregard the distinction between common and proper nouns. To this end, Table 3 reports the simple average over the two conditions; these are the kinds of averages that appeared in Figures 4 and 5 above. The probability of .73 in the top left corner now signals that, upon taking the simple average over common nouns (.59) and proper nouns (.87), the predicted probability of *s*-genitives for animate instances in press writing is .73. We can aggregate further by considering ANIMACY as our only targeted variable. GENRE then becomes a peripheral predictor, which means that we will also average over its levels. Giving all genres equal weight, this leaves us with the probabilities listed at the bottom of Table 3.

Table 3. Median predicted probability of *s*-genitives by ANIMACY and GENRE, averaged over NOUN TYPE.

	ANIMACY (averaged over NOUN TYPE†)		
GENRE	Animate	Collective	Inanimate
Press	.73	.53	.46
Adventure Fiction	.73	.22	.10
Non-fiction	.64	.28	.14
Learned	.67	.20	.11
General Fiction	.70	.10	.07
Averaged over GENRE†	.70	.27	.18

Note. †Simple (unweighted) average over all levels of the other predictor(s).

5.2. Simple vs. weighted averages

In our illustrative analysis, we have so far aggregated predictions using simple averages, assigning the same weight to all predictor levels. The predicted probability of .73 for animate referents in press writing, for instance, forecasts the share of *s*-genitives for a text assuming a balanced distribution of proper and common nouns. The observed overall distribution of noun types in the corpus, however, leans toward common nouns (65%). In fact, uneven distributions of cases across conditions are quite typical of natural language data, and there are settings where we may prefer to assign more weight to conditions that are more frequent in language use. We start with an illustrative comparison of different schemes and then consider the choice among them in light of the predictor classes discussed in Section 4.

To keep things simple, we will concentrate on the genre 'learned'. Let us assume that our interest centers on how the probability of *s*-genitives varies with the animacy of the possessor – ANIMACY is therefore our targeted variable. NOUN TYPE, on the other hand, is peripheral and we would like to average over its levels. We must then choose between a simple and a weighted average. We start by taking a closer look at the distribution of NOUN TYPE within each ANIMACY class. The percentages in Table 4 indicate that, in learned writing, possessors are generally more likely to be referenced by a common noun. This tendency is most pronounced for inanimate possessors (90% of the inanimate cases are referenced by a common head noun), followed by collective (71% common nouns) and animate possessors (58%). Note that we are now reporting percentages (instead of probabilities as in Tables 2 and 3). To (hopefully) avoid confusion, we use probabilities when reporting on the distribution of the *outcome* (i.e. the predicted probability of *s*-genitives) and percentages for describing the observed distribution of *predictor* variables in our dataset.

The observed distribution is shown in Figure 6a, where the area of the squares is proportional to the percentages listed in the table. The bottom row of Table 4 gives the overall distribution of NOUN TYPE in learned writing, disregarding ANIMACY subclasses. We will return to these percentages shortly.

Table 4. Distribution of NOUN TYPE levels by ANIMACY in learned writing.

	NOUN TYPE	
ANIMACY	Common	Proper
Animate	58%	42%
Collective	71%	29%
Inanimate	90%	10%
Overall distribution	78%	22%

Note. Data for the genre 'learned' only.

Note that both predictors are internal variables, i.e. outside the scope of the corpus compiler. They reflect natural distributional patterns in language use, and the overall rate of common and proper nouns makes sense linguistically. However, if we characterize ANIMACY levels by taking, for each category, a simple average over the predictions for common and proper nouns, as in Table 3 above, we are treating NOUN TYPE as though its distribution were perfectly balanced in language use. This weighting scheme, with uniform weights for common and proper nouns, is shown graphically in Figure 6b. We could instead opt for weighted averages that reflect the overall distribution of NOUN TYPE in the data (here: learned writing). Common nouns would then receive more emphasis (78%), which adds realism to the predicted probabilities. This set of weights is also shown graphically in Figure 6b.

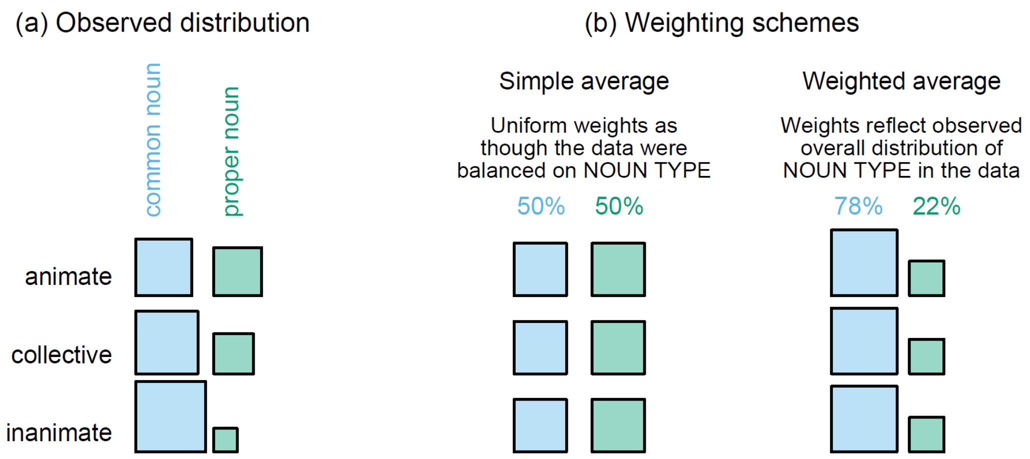

Figure 6. Observed distribution of noun types within each animacy class (panel a) and different weighting schemes for averaging over the peripheral factor NOUN TYPE (panel b).

Note that both weighting schemes disregard the differences between ANIMACY classes in the proportional share of proper vs. common nouns, which are evident from panel (a). Both sets of weights therefore allow ANIMACY levels to be compared on an equivalent basis, since predictions are adjusted for the association between ANIMACY and NOUN TYPE. For purposes of inference, this is often desirable, since it allows us to consider the net differences between animacy classes, setting aside variation that may actually be linked to another predictor, i.e. NOUN TYPE.

In principle, we could also use as weights the percentages reflected in Figure 6a, thereby preserving the association between the targeted and the peripheral factor. The resulting predictions will then carry a different meaning, however, since they conflate two patterns of variation: Differences among animacy levels and differences between noun types. Even though blended patterns of this kind may be of interest in certain research areas, we will not pursue this weighting strategy further here.

Let us now turn to a comparison of the two weighting schemes for our current set of data. Table 5 repeats, at the left end, the predictions for the full breakdown by ANIMACY and NOUN TYPE. For animate and inanimate possessors, we discern a difference between common (C) and proper (P) nouns. We now compare two ways of averaging over the factor NOUN TYPE. Using a simple average, we obtain a predicted probability of .67 for animate possessors, and .10 for inanimate ones – a difference of .57. If we adjust weights to preserve the overall distribution of proper and common head nouns in the learned section of the Brown corpus, the predicted probability for animate referents drops to .56 and that of inanimate referents to .06 – a difference of .50. In the present case, the choice of weighting policy matters for the comparison among animacy levels.

Table 5. Comparison of different weighting schemes for averaging over the levels of NOUN TYPE (C = common, P = proper).

	Conditional predictions		Predictions averaged over NOUN TYPE			
			Simple average		Weighted average	
ANIMACY	C	P	Weights (C, P)	**Pr**	Weights (C, P)	**Pr**
Animate	.48	.87	50%, 50%	**.67**	78%, 22%	**.56**
Collective	.21	.20	50%, 50%	**.19**	78%, 22%	**.20**
Inanimate	.04	.17	50%, 50%	**.10**	78%, 22%	**.06**

Note. Data for the genre ‘learned’ only.

Let us also consider this discrepancy graphically. In Figure 7, ANIMACY levels sit on the horizontal axis and the colored profiles denote the conditional predictions – proper nouns in green, common nouns in blue. The simple averages, which appear as a dashed grey profile, lie midway between these. The weighted average, on the other hand, which is shown in black, leans more heavily toward the prediction for common nouns, which outnumber proper nouns by a factor of about 4 to 1.

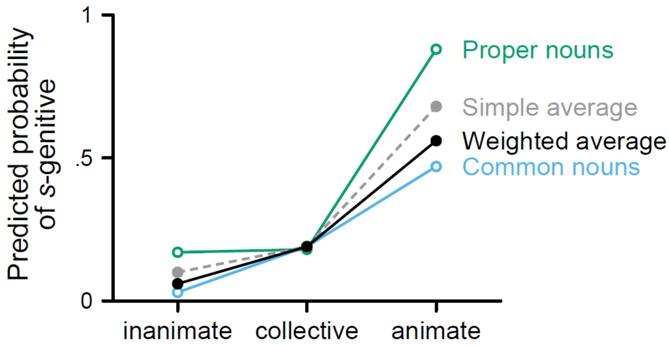

Figure 7. Comparison of different weighting schemes for averaging over the levels of NOUN TYPE.

When reporting predictive margins, we need to be clear about how the levels of peripheral factors were combined. The two sets of aggregated predictions in Table 6 and Figure 7 would be reported as averaged over NOUN TYPE, using either simple averages (first set) or weighted averages (second set). When weighted averages are used, we also need to state how the weights were constructed. In the present case, they reflect the observed overall distribution of the peripheral variable (NOUN TYPE) in the data.

5.3. Choice of weighting scheme

In situations where we are dealing with a larger set of predictor variables, some guidance can be provided on weighting policies for peripheral variables. This brings us back to the distinction between external and internal predictors (see Section 4).

Let us first consider the treatment of external peripheral variables. Since their distribution is usually not linguistically meaningful and will likely vary from corpus to corpus, it appears sensible to adjust predictions to represent a balanced scenario. Thus, for purposes of comparability, both within and across studies, we will usually opt for simple averages. To fix ideas, consider GENRE as a peripheral predictor. In our current set of data, the non-fiction section yields by far the greatest share of genitives (38%), the other genres hovering between 10% and 20%. Let's assume our targeted variable is ANIMACY, and we wish to average predictions over GENRE. Should we form simple averages and treat genres on an equal footing, or should we opt for weighted averages and thereby assign more emphasis to non-fiction writing? Simple averages appear to be the more sensible choice. In fact, we would argue for uniform weights as a default when averaging over external peripheral variables, especially when (i) the uneven distribution across their levels attracts no linguistic interest, and (ii) our corpus has not been compiled with reference to a meaningful sampling frame. A balanced weighting scheme also facilitates comparisons across corpora, where we should be adjusting predictions for differences in genre composition.

For internal predictors, on the other hand, the distributional patterns in our corpus reflect the inner workings of language (use). We would therefore routinely incorporate them into our predictions by using weighted averages. In situations where there are multiple internal peripheral variables, there are in fact different possibilities for constructing weights. First, each internal peripheral factor may be considered in isolation. The weights for its levels then mirror its *overall* distribution in the data. We will refer to this weighting scheme as *isolated weighting*. Put another way, *isolated weighting* uses the expected proportion of a given predictor level based on the marginal distribution of the predictor in the data. Alternatively, the set of weights can be modified to reflect the joint distribution of several (or all) internal peripheral variables in the model. We refer to the extreme case, where weights mirror the observed distribution among all internal peripheral variables as *joint weighting*. Of course, there are intermediate cases, which will not concern us here. The choice between these two extremes (i.e. between isolated and joint weighting) is far from trivial, and needs to be addressed in future research. Our tentative preference is to treat peripheral variables in isolation. The analyses in Section 7, however, apply both weighting schemes.

5.4. Handling quantitative variables

In contrast to the straightforward case of categorical predictors, quantitative variables require special treatment. While a random forest is run on the original, continuous representation of such predictors, model summaries may not be able to retain the existing level of detail. This is because forming predictive margins based on all possible values of a continuous predictor is often computationally prohibitive, especially in settings with multiple input features. We may therefore need to restrict our focus to a smaller set of values for which we then generate predictions.

This series of values should represent the range over which the predictor varies, and the number (and location) of points selected should strike a balance between (i) providing sufficiently detailed coverage for the detection of local non-linearities; and (ii) computational expense. We are hesitant to offer recommendations here, but it would seem that about 20 points may perhaps be a sensible choice.

A first step is to inspect the distribution of the predictor visually to determine an informative range of values. Note that this envelope need not extend to the smallest and largest value in the sample, especially if these lie far away from the bulk of the observations. Once a range has been determined, we could apportion it using an evenly-spaced set of bins. Another sensible approach would be to use quantiles of the distribution, to the effect that each point represents roughly the same share of cases.[4] The application of these general schemes is illustrated in the R tutorial (https://osf.io/r5uca). Note that the choice of how to bin continuous predictors will depend on the data and research aims. Thus, with the distribution of the two quantitative predictors in our model (THEMATICITY and LENGTH DIFFERENCE) being partially discrete, we opted for custom binning schemes, which we outline in web appendix 1 (https://osf.io/fz24j).

[4] The approach we take to quantitative predictors will run into problems if a random forest generates averaged predictions that fluctuate considerably at this level of resolution, i.e. in the direct neighborhood of the discrete points selected. This is because minor shifts in the location of these points may then alter the predictions and generate different patterns. To assess whether predictions show excessive micro-level wiggliness, partial dependence plots can be used as a screening device. Alternatively, if highly local oscillations of this kind seem implausible from a linguistic viewpoint, a viable strategy would also be to discretize the variable before fitting a random forest and then use the midpoints of the bins as a quantitative predictor.

Once a quantitative predictor has been discretized and appropriate weights constructed, it is treated in the same way as a categorical predictor. This completes our outline of the methodology for constructing predictive margins from random forests. Before we turn to a more comprehensive analysis of the genitive data, we contrast our approach with the related technique of partial dependence plots.

6. Comparison with partial dependence plots

Partial dependence plots (Friedman 2001) are constructed in a similar way to predictive margins: Predictions are obtained by feeding purposefully modified data into a random forest, and the predicted scores are then summarized. Modifications are made to the targeted variables, which are allowed to vary over the range of values found in the data. The treatment of peripheral factors differs, however. The predictions that appear in partial dependence plots are not built from reference grids and condition-specific weights. Rather, they are constructed from the underlying set of data.

To form partial dependence plots for the targeted variable ANIMACY, for instance, three copies are made of the original data table – one for each level of the targeted predictor. Then, the first copy is modified by setting all cases to “animate”. Predicted probabilities are then generated for this modified set of observations – one from each tree in the forest. The same is done for the other levels of ANIMACY: In copy 2, all observations are set to “collective”; in copy 3 to “inanimate”. This produces three distributions of predictions, which are then summarized and compared.[5]

From the viewpoint of our approach to forming predictive margins, partial dependence plots can be considered a special case. This means that equivalent summaries can be produced with the current approach, i.e. using a reference grid and condition-specific weights. To this end, we start by creating a grid with all possible combinations of predictor levels and values. For quantitative variables, the series of values must provide sufficiently detailed coverage. We then construct weights to reflect the joint distribution of *all* peripheral factors (i.e. internal *and* external ones) in the sample. The set of weights then essentially mimics the distribution of the cases in the original data table. With increased resolution of quantitative predictors, then, our predictive margins coincide with the patterns that emerge in partial dependence plots.

In contrast to partial dependence plots, which are restricted to one particular weighting scheme, our approach to forming predictive margins offers more flexibility. Thus, through our choice of weights, we can adjust predictions to improve comparability within and between corpora, or to correct for a lack of representativeness of our data with respect to a particular target population. Further, there is no upper limit on the number of targeted variables when forming predictions. Since random forests are given credit for their ability to capture complex interactions among predictors, the fact that current implementations of partial dependence plots are limited to two targeted variables may seem disappointing.

[5] The same approach to obtaining predicted probabilities is used when forming, from generalized linear regression models, what are referred to as “average marginal effects” (see Long 1997: 74), “average partial effects” (Wooldridge 2016: 179), or “average predictive comparisons” (Gelman et al. 2021: 249).

Finally, having a reference grid with the predicted probabilities of interest allows us to compute secondary quantities such as differences (or contrasts) between predictions. The ability to express specific comparisons directly can greatly facilitate interpretation. Our case studies, to which we now turn, should make the benefits derived from the added flexibility of predictive margins apparent.

7. Application to the English genitive alternation

Before we conduct a more comprehensive analysis of the genitive data, we sort variables into internal and external features. We then consider two analysis settings. The first represents a planned analysis, where key comparisons are formulated before the data are analyzed. The second analysis is an unplanned, exploratory search for patterns in the forest.[6]

7.1. Predictor classes and weighting scheme

Table 6 describes the set of predictors that enter the analyses and groups them into external and internal features. This classification is fairly straightforward.

Table 6. Overview and classification of variables in the analysis

Predictor	Description
External	
GENRE	Learned, press, non-fiction, adventure fiction, general fiction
TIME	1960s (Brown corpus), 1990s (Frown corpus)
Internal	
ANIMACY	Animate, collective, inanimate
NOUN TYPE	Common noun, proper noun
FINAL SIBILANT	Possessor NP ends with a sibilant: Yes/no
LENGTH DIFFERENCE	Possessor-to-possessum length ratio (based on word count)
GIVENNESS	Discourse status of possessor: New/given
PROTOTYPICALITY	Prototypical genitive relation (e.g. part-whole, kinship): Yes/no
THEMATICITY	Logged normalized text frequency of the possessor

When averaging over peripheral variables, we will use simple averages for external factors (GENRE, TIME) and weighted averages for internal variables. For weighted averages, we implement two schemes: isolated weighting and joint weighting. Recall that isolated

[6] While these modes of inquiry are reminiscent of the distinction between exploratory and confirmatory research (e.g. Tukey 1980), three points should be noted. First, confirmatory studies usually rely on purpose-built designs that employ randomization as a methodological safeguard. Experimental set-ups of this kind contrast starkly with the observational nature of corpus data, and it could be argued that genuinely confirmatory research should not rely on corpus data. Second, in carefully planned confirmatory studies, there is usually a straightforward relation between design and analysis, and data are better handled by ANOVA- or regression-style analyses. Finally, it is not at all clear whether pure prediction algorithms such as random forests have the capacity to produce valid statistical inferences (e.g. Efron 2020: 643).

weights only capture the overall, univariate distribution of each peripheral internal variable in the data. Joint weights, on the other hand, reflect the multi-way distribution of *all* internal peripheral variables in the data, which brings us closer to the implicit weighting approach implemented in partial dependence plots. The predictions shown in the following graphs rest on the isolated weighting scheme. For comparison, the joint-weighting counterparts of the following charts appear in web appendix 2 (https://osf.io/nujp9). We found no notable differences in the predicted patterns, apart from slightly wider uncertainty intervals for the isolated scheme.

7.2. Planned analysis: Time trends

The first analysis focuses on time trends, the key question being whether there is evidence for diachronic change in the genitive alternation. With comparisons between the Brown and Frown corpus taking center stage, TIME will always be a targeted variable. We will direct our attention to the consistency of diachronic patterns across conditions, that is, we will pair TIME with other targeted variables. If similar trends emerge in different subsets of the data, this will strengthen the case for change over time.

We will now outline different kinds of predictive margins we may form to investigate the presence and consistency of time trends. For a point of reference, we start with a global assessment, i.e. a comparison of the two points in time, averaging over all other predictors in the data. For the 1960s, this yields a predicted probability of .36 for the *s*-genitive, 90% PI [.32; .38], compared to .41 [.39; .44] for the 1990s – a difference of .05. The overall increase of the *s*-genitive therefore amounts to 5 percentage points.

To assess the stability of this trend, we consider each of the remaining predictors, in turn, as a secondary targeted variable. This reflects our data-analytic angle: We are asking whether time trends (primary) are stable across the levels or values of another (secondary) predictor. Accordingly, graphs put TIME on the x-axis, showing different profiles according to the secondary targeted feature. Figure 8 shows the profiles we obtain. Note that, for this purpose, the two quantitative predictors (LENGTH DIFFERENCE, THEMATICITY) have been coarsely categorized; the cutpoints are marked with grey triangles in Figures 1 and 2 of web appendix 1. Judging from Figure 8, the slopes are in good agreement, which indicates that time trends appear to be quite stable across these conditions.

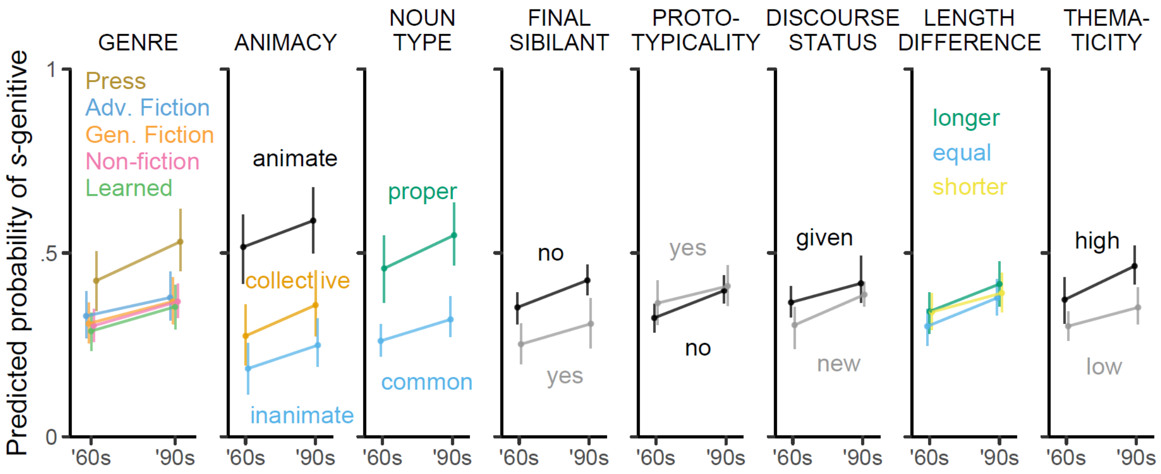

Figure 8. Predictive margins for TIME paired with each of the other predictors in the model. Error bars reflect 90% PIs. 🅭🅯

We can go further and form predictive margins for TIME with multiple secondary targeted variables. The visualization task then becomes more complex. While TIME, our primary targeted variable, remains on the x-axis, we need to resort to multi-panel arrangements using facetting. Figure 9 shows such an arrangement for ANIMACY and GENRE as secondary targeted features. Again, we get the impression that trends are coherent across conditions. Note that the predictive margins in Figure 9 provide a three-way breakdown of predicted probabilities (i.e. three targeted variables are considered at the same time). We could go on and investigate other combinations of targeted variables.

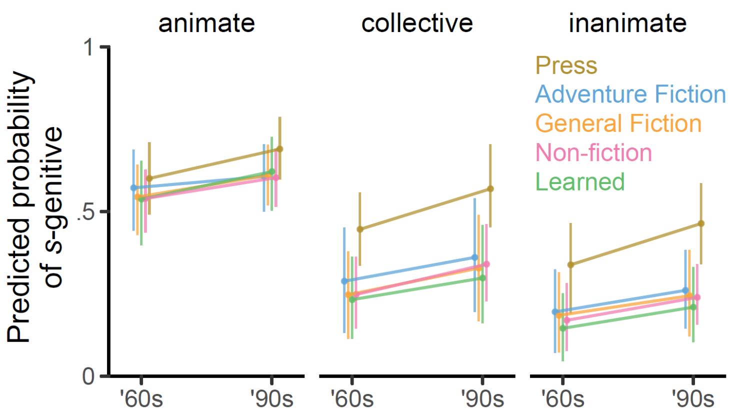

Figure 9. Predictive margins for TIME paired with two other targeted factors (ANIMACY, GENRE). Error bars mark 90% PIs. 🅭🅯

We conclude this first analysis of the genitive data by illustrating another data-analytic strategy, which capitalizes on the fact that (i) predictive margins can be formed for arbitrarily fine-grained multiway cross-classifications and (ii) the profiles of interest can be summarized by a secondary quantity, i.e. a difference between predicted probabilities. We refer to differences between predictive margins as *predictive contrasts*. Recasting diachronic patterns into predictive contrasts simplifies graphical displays considerably. As a result, we can study a much more nuanced array of conditions. We carry this task to the limit by considering all possible cross-classifications of the categories that appear in Figure 8 above. Going from left (GENRE) to right (dichotomized THEMATICITY) we obtain a total of 5 x 3 x 2 x 2 x 2 x 2 x 3 x 2 = 1,440 conditions. We cannot obtain predictions for this full set, since not all of these combinations are attested in the data – for 13% of the conditions, our forest fails to generate a prediction. This leaves us with 1,250 unique attested combinations. The predicted time trends for these appear in Figure 10a. The majority of slopes point upwards, which is confirmed by the distribution of predictive contrasts. These are shown in panel (b). Differences range from roughly −.10 to +.30; 11% are negative, and 83% positive (6% are 0). We can conclude that, in general, the directionality of time trends remains consistent as we zoom into this more comprehensive set of conditions. Diachronic patterns therefore seem to have considerable generality across the conditions studied.

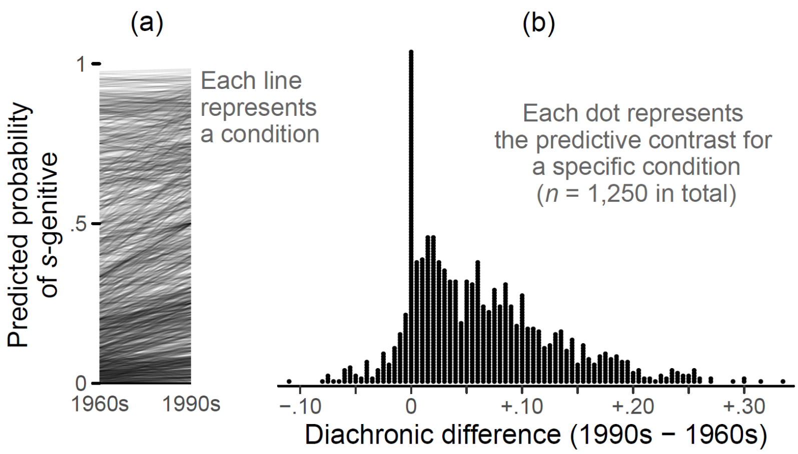

Figure 10. Predictive margins for TIME: (a) trends for 1,250 conditions; (b) predictive contrasts for the same set of conditions.

7.3. Post-hoc exploration

Our second analysis sets out with no specific priorities. Rather, our interest is in the kinds of patterns documented by our random forest. We can nevertheless distinguish a number of exploratory directions:

- For each factor, we would like to observe how the predicted probability of *s*-genitives varies across its levels or values, and whether this arrangement is consistent with previous research.
- For quantitative variables, we attend to the shape of the association, asking whether a straight line offers an acceptable summary, or whether notable non-linearities produce a more complex form of relationship.

- Interest often also attaches to the relative importance of predictors, and we may wish to rank them in terms of their strength of association with the outcome.
- A final concern is with the presence and form of interaction patterns among predictors.

We start by forming predictive margins for each predictor in turn. The patterns we obtain appear in Figure 11. For categorical inputs, we have ordered levels based on their predicted share of *s*-genitives. We note that press sticks out (relative to the other genres); the same is true for animate possessors (relative to inanimate and collective ones). For LENGTH DIFFERENCE, the predicted pattern is quite wiggly, with a notable peak at a length ratio of 2 to 1, i.e. cases where the possessor is twice as long as the possessum. For THEMATICITY, on the other hand, a straight line offers a good summary of the observed association.

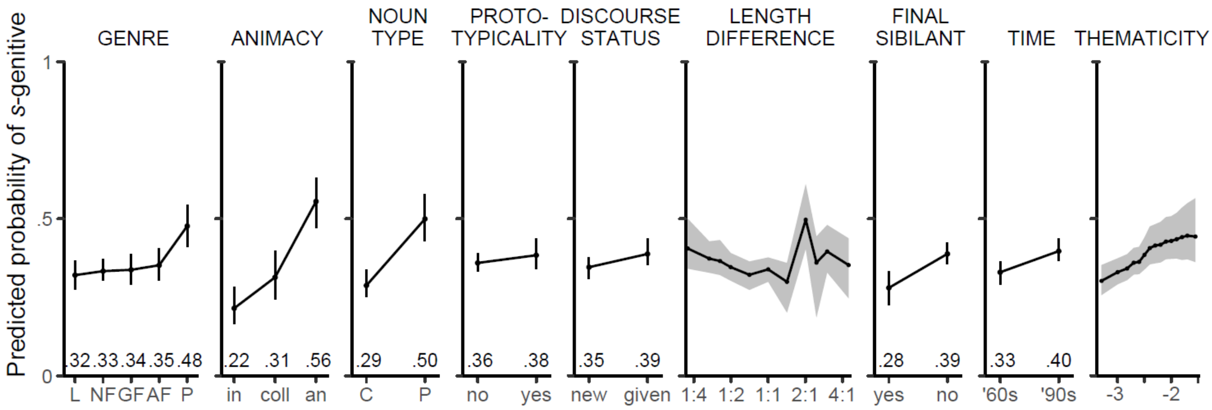

Figure 11. Predictive margins for each factor. Error bars/bands denote 90% PIs.

Next, we consider the relative importance of variables. The magnitude of the differences among predicted probabilities provides one assessment. To this end, we visually scan the vertical range of predictions within each panel in Figure 11. This produces the following approximate rank order: ANIMACY > NOUN TYPE > GENRE > THEMATICITY > FINAL SIBILANT, LENGTH DIFFERENCE > TIME, DISCOURSE STATUS > PROTOTYPICALITY. This rank order is in fairly good agreement with the standard variable importance scores for the random forest. For comparison, these are shown in Figure 12.

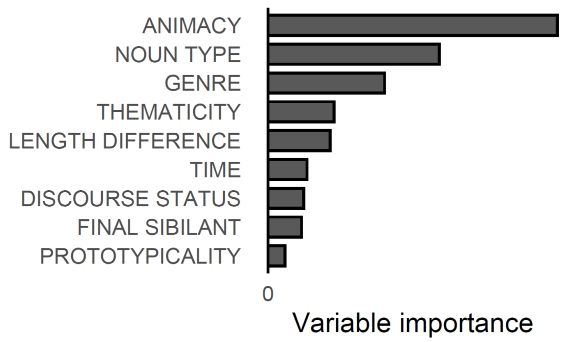

Figure 12. Variable importance scores for the genitive random forest.

Our final step is to search the forest for interactions among predictor variables. With a large number of inputs, it is quite difficult to know where to start. We can rely on two aids. First, it is often the case that relatively important predictors are more likely to also partake in non-negligible interactions (Gelman et al. 2021: 136). This means that we would direct our search towards those variables that rank high in terms of predictor importance, i.e. ANIMACY, NOUN TYPE, and GENRE. Assistance is also provided by ML-based indexes of interaction strength, such as Friedman's H (Friedman and Popescu 2008). These provide an indication of the degree to which a predictor is involved in interactions with other factors (see Grafmiller 2023). The H-statistics in the current forest also point to ANIMACY, NOUN TYPE, and GENRE. We therefore concentrate on these variables.

We directly turn to a consideration of the three-way arrangement, i.e. predictive margins for all 30 combinations of levels. Graphically, these multi-way comparisons can be arranged in different ways, since each predictor may be shown (i) on the x-axis, (ii) as a color-coded grouping variable, or (iii) as a facetting variable with subgroups assigned to different panels. When working from an exploratory perspective, we recommend considering different perspectives. Recall, however, that, in terms of interpretation, the x-axis hosts the primary peripheral variable, which is reflected in how a comparative query is phrased. Thus, Figure 13 corresponds to the following question: Do animacy profiles (primary) vary across noun types, and are these two-way configurations similar across genres?

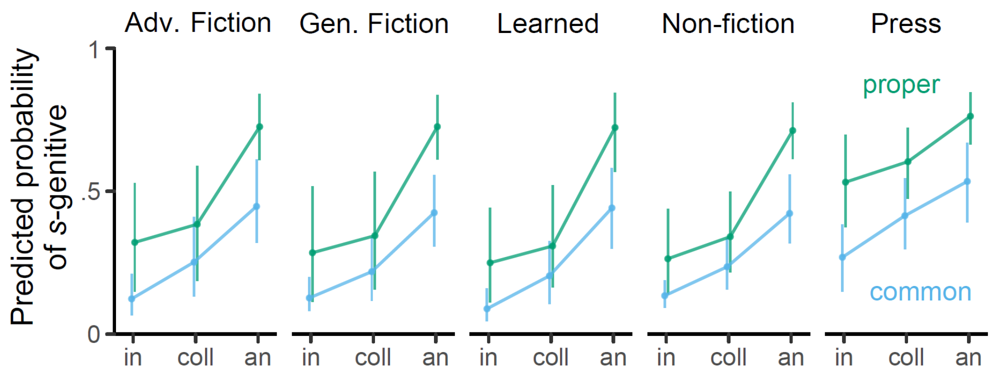

Figure 13. Predictive margins for three targeted variables (ANIMACY, NOUN TYPE, GENRE): Arrangement 1. Error bars: 90% PIs.

Figure 14 reorganizes these predictive margins and maps onto a slightly different query: Do differences between proper and common nouns (primary) vary across animacy levels, and are these two-way configurations similar across genres? Depending on our research focus and the focal comparisons, we may also try other arrangements, with GENRE on the x-axis or as a grouping variable.

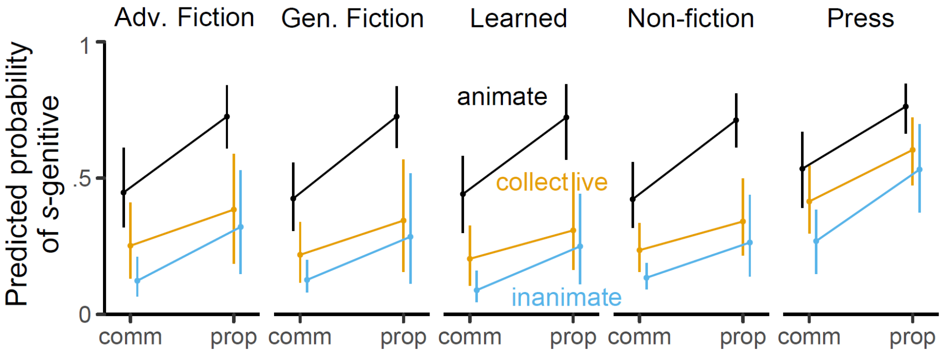

Figure 14. Predictive margins for three targeted variables (ANIMACY, NOUN TYPE, GENRE): Arrangement 2. Error bars: 90% PIs.

8. Summary and outlook

Tree-based methods are by now firmly integrated into the statistical toolkit of corpus linguistics, not least because they are capable of handling difficult data settings. The predictive perspective instantiated in these procedures may not comply with our research priorities, however. Thus, if the object of the analysis is to appreciate and understand the patterns formed by predictor variables, the standard output of tree-based methods (trees and variable importance scores) usually provides us with insufficient information. The present paper has illustrated the use of predictive margins as a general framework for gaining instructive insights into a random forest model. We laid out the general procedure, which involves setting up a reference grid listing all conditions of interest, and then generating predictions for these from the collection of trees in the forest. In terms of general strategy, we discussed the treatment of quantitative variables and provided some linguistic orientation for the question of how to summarize predictions, in particular how to average over peripheral features. We adopted the distinction between internal and external predictors to suggest sensible weighting schemes when averaging over the levels of non-focal inputs.

Once the reference grid is set up, we can tailor the weighting scheme to our linguistic objectives. This part of the procedure is rather intricate, and we acknowledge that, in suggesting some principled ways of forming weights, we have merely scratched the surface of this methodological issue. We believe, however, that the distinction between internal and external features should prove useful for future work. In particular, the proposition of preserving, in a set of adjusted predictions, the distributional profiles of predictors that capture the inner workings of language seems to be worthy of more general consideration in language data analysis.[7] Further, enhancing corpus comparability by adjusting for imbalanced external predictors also appears to be a worthwhile avenue for methodological work. The fact that, in our case study, the choice of weighting scheme (i.e. isolated vs. joint

[7] Note that the issue of weighting is just as relevant for the interpretation of regression models based on observational data, a topic that does not seem to be discussed often in the (statistical) literature.

weights for internal peripheral factors) did not produce noticeable shifts in the observed patterns should not lead us to assume that it is generally irrelevant. Rather, it appears to be inconsequential if there are no appreciable interactions between the targeted predictor(s) and the set of internal peripheral variables. More research is needed, however, to understand the role and linguistic significance of weighting schemes when forming predictive margins (not only) from random forests.

The prediction grid we obtain by appending to our reference grid the full set of forecasts from the random forest grants us substantial flexibility when forming predictive summaries. As we have illustrated, there are no constraints on the combinatorial possibilities, which allows us to look for higher-order interactions in the forest and appreciate the degree of complexity implied by the model. We have seen that line plots are suitable vehicles for graphing multi-way predictor arrangements. Used in combination with faceting, they can easily host three targeted variables, which makes them an ideal companion for the visualization of predictive margins. We have also demonstrated the use of predictive contrasts to simplify a complex array of predictions and make it amenable to graphical display. Thus, by representing the comparison between predictions for the 1990s and 1960s as a simple difference we were able to assess the generality of diachronic trends across a fine-grained set of conditions.

The aim of the present paper has been to illustrate how to obtain predictor patterns from a random forest. We must concede that this technical focus has diverted our attention from questions of model adequacy, which are also relevant for ML procedures. While we have tacitly assumed that our data sit well with the assumptions underlying a random forest model, there is structure in our set of observations that cannot be accommodated by the algorithm we have implemented here. Thus, due to the design of our corpora and the nature of the grammatical phenomenon we are studying, our data points are in fact structured along two additional dimensions, which we have failed to communicate to our model. Specifically, genitive tokens are clustered (i) by source, since each text file contributes multiple data points to our study; and (ii) on linguistic grounds, since we may count multiple instances of the same possessor head noun (e.g. *city*, *President*). In regression parlance, the genitive data therefore show a (partially) crossed random-effects structure. Since these sources of variation can have noticeable effects on the shape of predictor patterns and their statistical uncertainty, they need to be incorporated into a model if it is to provide valid descriptions of and inferences from data. Seeing that the resulting non-independence of observations is a typical feature of corpus data and language data more generally (see Winter and Grice 2021), we urge researchers to (routinely) give due consideration to questions of model adequacy.

We nevertheless hope to have illustrated the utility of predictive margins for constructing informative and accessible summaries of random forest models. Since (corpus) data analysis often centers on predictor variables and the patterns they form, this framework gives us an interpretative handle on predictive models and allows us to see the wood for the trees.

References

Breiman, Leo. 2001. Random forests. *Machine Learning* 45(1), 5–32. doi:10.1023/A:1010933404324

Breiman, Leo, Jerome H. Friedman, Richard A. Olson & Charles J. Stone. 1984. *Classification and regression trees*. New York: Chapman & Hall.

Efron, Bradley. 2020. Prediction, estimation, and attribution. *Journal of the American Statistical Association* 115(530), 636-655. doi:10.1080/01621459.2020.1762613

Francis, W. Nelson & Henry Kučera. 1964. *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English for use with Digital Computers*. Providence, Rhode Island: Department of Linguistics, Brown University. http://icame.uib.no/brown/bcm.html (accessed 22 February 2023). .

Friedman, Jerome H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29(5), 1189–1232. doi:10.1214/aos/1013203451

Friedman, Jerome H. & Bogdan E. Popescu. 2008. Predictive learning via rule ensembles. *The Annals of Applied Statistics* 2(3), 916–954. doi:10.1214/07-AOAS148

Gelman, Andrew, Jennifer Hill & Aki Vehtari. 2021. *Regression and other stories.* Cambridge: Cambridge University Press.

Grafmiller, Jason. 2023. The genitive alternation in 1960s and 1990s American English: Data from the Brown and Frown corpora, https://doi.org/10.18710/R7HM8J, DataverseNO, V1.

Grafmiller, Jason. 2023. Visualizing grammatical similarities in comparative variationist analysis. In Lukas Sönning & Ole Schützler (eds.), *Data visualization in corpus linguistics: Critical reflections and future directions*. Helsinki: VARIENG. http://www.helsinki.fi/varieng/journal/volumes/22/grafmiller/.

Gries, Stefan Th. 2020. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory* 16(3), 617–647. doi:10.1515/cllt-2018-0078

Grissom, Robert J. and John J. Kim. 2012. *Effect sizes for research: Univariate and multivariate applications*. New York: Routledge.

Hundt, Marianne, Andrea Sand & Paul Skandera. 1999. *Manual of information to accompany the Freiburg–Brown Corpus of American English* ("Frown"). Freiburg: Englisches Seminar, Albert-Ludwigs-Universität Freiburg. http://korpus.uib.no/icame/manuals/FROWN/INDEX.HTM (accessed 22 February 2023).

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Lane, Peter W. & John A. Nelder. 1982. Analysis of covariance and standardization as instances of prediction. *Biometrics* 38(3), 613–621. doi:10.2307/2530043

Levshina, Natalia. 2020. Conditional inference trees and random forests. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*, 611–643. New York: Springer. doi:10.1007/978-3-030-46216-1_25

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks, CA: Sage.

Molnar, Christoph. 2019. *Interpretable machine learning. A Guide for making black box models explainable.* https://christophm.github.io/interpretable-ml-book/.

R Core Team. 2022. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Shah, Priti, & Eric G. Freedman. 2011. Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Science* 3(3), 560–578. doi:10.1111/j.1756-8765.2009.01066.x .

Sönning, Lukas. 2023. Drawing on principles of perception: The line plot. In Lukas Sönning & Ole Schützler (eds.), *Data visualization in corpus linguistics: Critical reflections and future directions*. Helsinki: VARIENG. http://www.helsinki.fi/varieng/journal/volumes/22/soenning/.

Strobl, Carolin, James Malley & Gerhard Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14(4), 323–348. doi:10.1037/a0016973

Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller & Melanie Röthlisberger. 2016. Around the world in three alternations: Modeling syntactic variation in varieties of English. *English World-Wide* 37(2), 109–137. doi:10.13140/RG.2.1.1886.0248

Tagliamonte, Sali & R. Harald Baayen. 2012. Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24(2), 135–178. doi:10.1017/S0954394512000129

Tukey, John W. 1980. We need both exploratory and confirmatory. *The American Statistician* 34(1), 23-25. doi:10.2307/2682991

Winter, Bodo & Martine Grice. 2021. Independence and generalizability in linguistics. *Linguistics* 59(5), 1251–1277. doi:10.1515/ling-2019-0049

Wright, Marvin N. & Andreas Ziegler. 2017. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* 77(1), 1-17. doi:10.18637/jss.v077.i01

Wooldridge, Jeffrey M. 2016. *Introductory econometrics: A modern approach*. Boston, MA: Cengage Learning.

Yarkoni Tal & Jacob Westfall. 2017. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science* 12(6), 1100–1122. doi:10.1177/1745691617693393